Lecture Notes:
Some Core Ideas of Imputation for Nonresponse in Surveys

Tom Rosenström
University of Helsinki

May 14, 2014

Contents

1 Preface

These lecture notes briefly describe some of the core ideas in the widely applied imputation methods for nonresponse in survey datasets. It is assumed that the reader has heard about possible advantages that can result from imputation of missing data, but wishes to gain a better intuition about how and why these methods work. The aim is to provide background knowledge that helps reading practical guides published in applied-research literature, such as the journals Statistics in Medicine (White, Royston, & Wood, 2011) and Psychological Methods (Schafer & Graham, 2002). The focus is on conceptual understanding, not on any particular statistical software. The concepts are illuminated through simplified simulated datasets, because then the true data generating mechanisms are known, and therefore also the relative success of the methods. These notes were put together rather fast, and I may improve clarity and accuracy of details and add material later on if the notes turn out useful for others. Lots of published material exists on the topic, but one should keep in mind that the imputation research is a relatively new field of study and many questions still remain without a good answer.

2 Definitions

Donald Rubins book is a classic reference on imputation methods (Rubin, 1987). Rubin introduced a way to classify missing data that is now in common use. We will use that classification too, and other common notation from basic statistics and probability theory. For example, capital letter Y denotes a random variable, and lower case letter y its realized value. Then $P(Y = y)$ is a probability of the event that the variable Y happens to get the value y. $P(Y)$ is a shorthand for distribution or arbitrary event on Y, and $p(y)$ for the density function.

It often happens in survey studies that all participants have not reported their value of Y, but we still think that such a value exists for those people too. In order to further dicuss about such cases, we denote those unobserved but existing values by Y_{mis} and the actually observed values by Y_{obs}. Complete data consists from observed and unobserved values, $Y_{com} = (Y_{obs}, Y_{mis})$. We also need to define a random variable R for the missingness status, so that for each observational unit, or participant,

$$R = \begin{cases} 1 & \text{if observation of } Y \text{ is missing} \\ 0 & \text{otherwise.} \end{cases}$$

Let us assume that we are interested on a phenomenon governed by the variable Y and have observed it for 1000 participants. For each participant, we throw a dice and erase the observed value y if we get two or less from the dice-throw variable Z. In this case, we know that missingness R is statistically independent of the variable of interest Y, although it depends on an irrelevant phenomenon captured by Z. We say that the data is *missing completely at random* (MCAR). Formally MCAR means that

$$P(R|Y) = P(R). \tag{1}$$

In other words, missingness does not depend on the variable(s) of interest. This is a good situation, because only thing we lose due to the strange dice-throwing exercise is sample size and statistical power. Unfortunately, it is rarely feasible to assume that R depends on such completely unrelated variable Z as a throw of a dice. It might occur, for example, that Y stands for a depression score and depressed participants stay home instead of showing up in our follow-up study. This situation could mean that Y is missing whenever, say, $Y > 1$. Clearly then, the above equation will not hold, and we say that the data is *missing not at random* (MNAR). In this case, we both lose power *and* get bias to statistical estimates based on Y_{obs} only. Figure 1 shows an example of a variable for which there appears to be selective attrition on the high values. Indeed, the average value of complete data is -0.01, whereas the average value of observed data is -0.60. If we wish to estimate the true population mean with the average, then the difference 0.59 represents a bias in our estimate. Using average of Y_{obs}, we underestimate the average amount of depression in the population.

In real data-analysis problems, we do not have access to y_{com} but only to y_{obs}, and we typically lack knowledge about the mechnanisms of missingness. That is, we only observe the right histogram of the Figure 1, and can mostly speculate whether or not it is subject to systematic attrition[1]. It is a common occurrence, however, that for a participant with a missing value in Y, some other variable X is not missing. If one represents both the variables in a data frame $D = (X, Y)$, then it can happen that

$$P(R|D_{com}) = P(R|D_{obs}). \tag{2}$$

In this case, where the propabilities of missingness depend on observed data but not on missing data, the data is said to be *missing at random* (MAR), or missingness is said to be *ignorable* (when not ignorable, it is *nonignorable*). For MAR data, it is often possible to correct estimate bias using modern imputation methods, or missing-data models. So Rubin's "missing at random" means that *given the observed data* the missingness is random (or independent of the variables of interest), whereas "missing completely at random" means that missingness is random with respect to both observed and unobserved data.

In what follows, we will only concentrate on the MAR case, for which general missing-data models can be built. Typically, MAR is only an assumption that cannot be directly tested, except by obtaining follow-up data from the nonrespondents. In fact, we frequently expect some deviation from the MAR assumption, acknowledging that small deviations are not likely to degrade much the performance of the MAR-based methods. That is, we would be worse of by not using these methods than by using them: even if we do not achieve unbiased estimates, at least they are less biased.

3 Different ways to handle MAR data

Figure 1 in the previous section showed histograms for a simulated variable Y, and we will continue the discussion using the same simulated data with another variable, X, included. The below table shows few first observations of the full data frame of 1000 simulated observations. In this table, "NA" denotes

[1] If one is willing to make additional assumptions, it may became possible to correct bias in an estimate even in this case. If we have a reason to believe that Y_{com} is normally distributed, we can use knowledge about Normal distributions to adjust estimates based on Y_{obs}. Such assumption is rarely feasible, however.

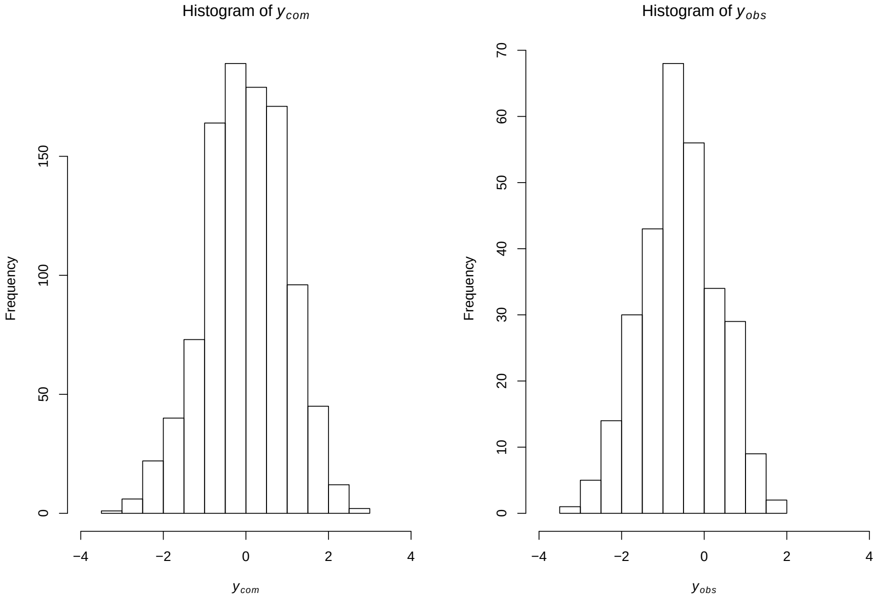
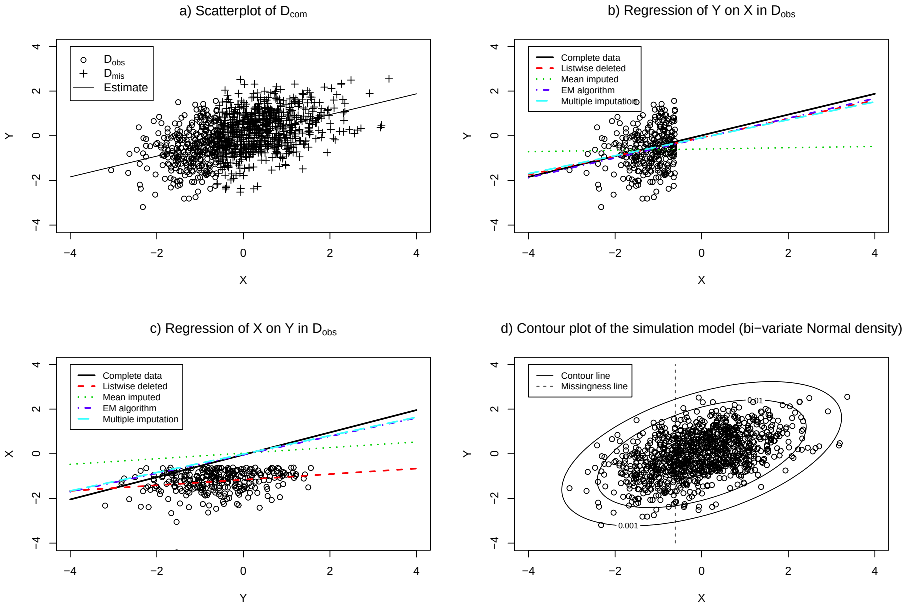


Figure 1: Histograms of complete data (left) and observed data (right)

a missing value. Altogether 709 Y-values out of the 1000 are missing. For simplicity, only two variables and the missingness-variable R are shown, but the discussion can be extended to arbitrary numbers of variables. A common way to handle missing values in Y is to remove all rows in the data frame that have "NA" and proceed with the data analysis by treating the cropped data frame as the only and complete information. This approach is known as *listwise deletion* of missing data. It is a good first step, but can lead to biased estimates and reduces statistical power. The bias can also extend to estimated relationships between the variables. For example, estimate of correlation in the listwise-deleted data frame is 0.23 with a 95% confidence interval of (0.12, 0.33). Because in this simulated example we have an access to the complete data frame, we can compute that without the missing observations the correlation would be 0.48 with a confidence interval of (0.43, 0.53). In this example, the bias in the correlation estimate is considerable (52% of the estimate). When estimating the population parameter, the listwise-deleted data would lead us to infer a small to medium relationships between X and Y, whereas the complete data suggests a large correlation.

X	Y	R
0.81	NA	1
0.04	NA	1
-1.12	-0.36	0
-0.64	-0.80	0
$\vdots$	$\vdots$	$\vdots$

Figure 2b shows what happens when we estimate a linear model from the data that was suhject to listwise deletion of partially missing observations (red, dashed line), and compares the regression estimate with that from the complete data case (black, solid line) and those using different imputation methods. Only one of the approaches produces a clearly biased estimate: mean imputation. ”Mean imputation” refers to replacing of Y_{mis} with the mean of Y_{obs}. It is easy to see why the mean imputation performs badly. Due to the positive linear relationship in the complete data, grand mean of Y is always below the conditional mean for participants with high values of X. Because the attrition occurs selectively in those with high values of X, substituting mean of Y in the place of their true Y value results in a downward shift compared to true complete data. Therefore, the regression slope estimated from the mean-imputed data is downwards biased compared to the complete-data estimate. Bias is not the only issue with mean imputation, however.



a) Scatterplot of D_{com}

b) Regression of Y on X in D_{obs}

c) Regression of X on Y in D_{obs}

d) Contour plot of the simulation model (bi−variate Normal density)

Figure 2: Results on simulated data. *a*) Complete data, with missing data points marked by the crosses and observed data points by the circles. The solid line shows a linear regression estimate. *b*) Different regression approaches for partially missing data. Complete-data estimate is shown for a reference (black, solid line). Other approaches are listwise deletion of partially missing observations (red, dashed line), replacing missing values with the variable mean (Mean imputation; green, dotted line), Expectation-Maximization (EM) algorithm (dark blue, dash-dotted line), and Multiple imputations of missing values from a probability model (light blue, dashed line with long dashes). *c*) Same as panel *b* but the estimates are for linear model with Y predicting X rather than other way around. *d*) The data was simulated from a bivariate Normal distribution with mean (0, 0) and variances of 1, and covariance of 0.5. Contour lines of the data distribution are shown by the solid lines, whereas the dashed line shows the quantile 0.25 of X. Values exceeding that threshold in X were set as missing in Y, creating a data set with missing values missing at random (MAR).

Standard error of the simple linear regression coefficient is of the form

$$S.E.(\hat{\beta}) \approx \frac{\hat{\sigma}}{s\sqrt{n}},$$

where $\hat{\sigma}$ is residual standard deviation, s standard deviation of the independent variable, and n stands for the sample size, or number of participants. Therefore, replacing all missing values of Y with one and the same value, mean of Y_{obs}, simultaneously both decreases $\hat{\sigma}$ (because identical values are easy to predict) and increases n relative to listwise deletion. This means that we both have a biased estimate *and* too much confidence in it (downwards biased estimate of error)! Hence, mean imputation is NEVER a wise strategy for handling missing data and nonresponse.

Figure 2b suggests that listwise deletion, EM-algorithm, and Multiple Imputation are not overly biased. This may come as a surprise for the reader, as listwise deletion is the simplest way to handle the partially missing data. Hence, listwise deletion can sometimes perform well, giving more or less the same estimates as more advanced imputation methods. Yet, should we try to predict X with Y rather than Y with X (a completely legitimate question as these are just symbols), listwise deletion is perhaps even more biased than the mean imputation (see Figure 2c). In this case, EM-algorithm and Multiple imputation are significantly better options than listwise deletion. Recall also that the correlation coefficient was attenuated in D_{obs} compared to D_{com}. So, the take-home message here is that listwise deletion sometimes performs well and other times badly, and because advanced methods introduce additional complexity and possible errors, it can be a good idea to check also the simple listwise-deletion estimates as a sensitivity analysis and additional information (White, Royston, & Wood, 2011). Exercise 1 asks the reader to further ponder why listwise deletion here leads to bias when X is regressed on Y, but not when Y is regressed on X.

Imputation methods that introduce biased estimates or their standard errors, such as mean imputation and listwise deletion, are called *improper* imputations. For example, replacing missing Y_{mis} with best linear estimate based on observed pairs (Y_{obs}, X_{obs}) would introduce less bias than mean imputation herein, but does not solve the variance-restriction issue and therefore leads to overconfidence in subsequent estimates. Methods that produce unbiased estimates and also properly handle their uncertainty estimation are called *proper* imputations. In the following sections, we will discuss about proper imputation methods. Several proper imputation methods exist, each taking a certain position in a continuum from ”classic” *versus* ”Bayesian” stance on statistics. In general, classical or frequentist position in statistics implies that such a thing as probability for regression coefficient does not exist. One cannot ask what is the probability that β exceeds zero, $P(\beta > 0)$. A frequentist modeler thinks that there exists just one ”true”, although unknown, population parameter β, and one can only make inferences about the probability of observing data under the model and given some value β. So only data are random variables, model characteristics being fixed thruths. In Bayesian statistics, however, also model parameters are thought as random variables and it is perfectly correct to ask for the probability $P(\beta > 0)$. Even if the Bayesian statistician would also think that only single ”true” β exists, he is willing to associate a probability with his degree of belief regarding that true value[2].

4 Maximum Likelihood methods

Both Frequentists and Bayesians study likelihood functions. They arise when a statistical probability model is given/assumed for the data. For example, I simulated the data used here from a bivariate Normal distribution, as independent observations. The density function of this distribution is

$$f((x,y);\mu,\Sigma) = \frac{1}{(2\pi)|\Sigma|^{1/2}} e^{-\frac{1}{2}((x,y)-\mu^T)\Sigma^{-1}((x,y)^T-\mu)}, \tag{3}$$

where $\mu^T = (\mu_x, \mu_y)$ is the mean vector of X and Y, and Σ their covariance matrix[3]. Here they were given values $\mu^T = (0,0)$ and

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

When we fix these parameters for simulation, $f((x,y);\mu,\Sigma)$ characterizes the probabilities/frequencies for getting a given value (x,y). But when we have observed values x_i and y_i for an individual i, $f((x_i,y_i);\mu,\Sigma)$ can be thought as a function of μ and Σ rather than (x_i,y_i). That is, now the observed values are fixed and f changes as a function of the Normal-distribution parameters. The function f is such that it grows when μ gets closer to observed values (x_i,y_i), and is called the *likelihood function* of the parameters μ and Σ. Maximum Likelihood methods use this observation. They maximize the likelihood function for the observed data in order to find the most likely parameters underlying the data[4].

The problem with listwise deletion is that we lose all the information that an observation of X would provide on $f((x,y);\mu,\Sigma)$ just because of the lacking y. And because of the correlation between X and Y, the observation x also provides information regarding the unobserved associated y. This is obvious from the contour plot in Figure 2d: for a high x, lower values of y are less likely than higher values. Expectation Maximization (EM) algorithm corrects this shortcoming. And in the above simulation example, this also corrected the bias in listwise-deletion estimate! This is because (1) the MAR assumption held and (2) because we used a correct likelihood function (i.e., correct statistical model). (1) After taking the value of X and the correlation of X and Y into account, missingness in Y is essentially random. (2) We had a correct model assumption, because we know that I simulated the data from that model. Hence,

[2] To make a complex issue simple, frequentist methods are often simpler to compute in practice and involve less flexible options and therefore less complexity than their Bayesian counterparts. On the other hand, each frequentist method can be derived from the Bayesian perspective too, and Bayesian methods are superior in most, or all, other aspects than simplicity. For example, a Bayesian can ask the direct questions that typically most interests the researcher, such as the probability $P(\beta > 0)$.

[3] Here the operation $\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}^T \mapsto (\mu_x, \mu_y)$ is a transpose of the column vector. It is given for notational consistency, and a reader not familiar with matrices need not worry about it.

[4] Because we assume that participants are statistically independent of each other, the likelihood of all the data is just a product of the likelihood functions for each participants: $L(\mu,\Sigma) = \prod_{i=1}^{n} f((x_i,y_i);\mu,\Sigma)$. Maximizing this product effectively summarizes information from the observervations to the population parameters μ and Σ

in order to use the EM algorithm, one must assume both MAR missingness and a statistical model (probability distribution) for the data. Both can go wrong. Slightly wrong assumption is often less bad than listwise deletion, but grossly wrong assumptions may lead to a worse end result.

We will not go into details of the EM estimation, but a brief explanation is in order. The EM algorithm is an iterative procedure for drawing information about missing values from the assumed model and using both the information and observed data to make inferences about that model (i.e., to fix its parameters). Let $L(\theta; D_{obs}, D_{mis})$ be the likelihood function of all data, D_{com}, given a parameter vector θ; when interpreted as a function of data, a data density $L(\theta; D_{obs}, D_{mis}) = p(D_{obs}, D_{mis}; \theta)$. In order to use EM algorithm, we need to be able to solve for the conditional data density of missing values given the observed ones and the parameters, to write down $p(D_{mis}|D_{obs}, \theta)$ so that we can compute expected values of functions under this probability distribution, $\mathrm{E}_{(D_{mis}|D_{obs},\theta)}[g(D_{mis})] = \int g(z)p(z|D_{obs}, \theta)dz$. EM algorithm takes an arbitrary value for the estimate of $\theta^{(0)}$ and repeatedly iterates the following two steps

1. **Expectation step (E step):** Compute the expectation

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{(D_{mis}|D_{obs},\theta^{(t)})}[\log L(\theta; D_{obs}, D_{mis})].$$

2. **Maximization step (M step):** Find the maximum of the resulting function of θ, and set it as the new parameter estimate

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

After sufficiently many iterations, successive estimates $\theta^{(t)}$ and $\theta^{(t+1)}$ no longer differ, and the algorithm has converged to a stable and correct estimate for the parameters (or rarely, to a sadle point or parameter-space boundary). A perceptive reader may have noticed that imputed values are implicit in EM algorithm. That is, one does not create explicit new data points to replace the nonresponses, but the conditional distribution of these random variables enters into the expected value of the ”E step”, ’guiding’ the final estimate towards feasible parameter values under the chosen model. If one can simulate random values from the density $p(D_{mis}|D_{obs}, \theta)$, one could create a large number of explicit imputed values and compute the E step by numeric Monte Carlo integration, but this is rarely done in practice. EM algorithm is a good method where the E and M steps can be computed analytically (with ’pen and paper’), but where the calculations get difficult, other approaches tend to be favored. In practice, EM algorithm is mostly used for multivariate normally distributed, or approximately normally distributed, data that have partially missing vectors among the observations. It is also used in many entirely different contexts, for example, when handling unobserved (i.e., ’missing’) cluster memberhip variables in Finite Mixture Models (McLachlan & Peel, 2000).

EM algorithm has been around for a long time, and much has been written about it. A standard reference is Dempster, Laird, & Rubin (1977). Because the EM algorithm operates only on the likelihood function, and other data densities, such as $p(D_{mis}|D_{obs}, \theta)$, treating model parameters as fixed unknown quantities, it is a classical Maximum Likelihood procedure. Also other Maximum Likelihood methods for missing data exist, such as Full Information Maximum Likelihood in the Structural Equation Modeling context (Muthen, Kaplan, & Hollis, 1987; Enders, 2001), but we do not discuss them here.

5 Multiple Imputation methods

In these lecture notes, we mainly get to know better a one more missing-data technique: Multiple Imputation with Chained Equations. Multiple Imputation is a flexible imputation method that is easier and more automatic to apply than full Bayesian analysis, and yet allows proper imputations for a wide range of models. It has the advantage of supporting many of the widely applied classical/frequentist statistical methods, such as generalized linear models. Some Bayesian ideas are used in Multiple Imputation, however, and with almost the same effort we get to peek at the fully Bayesian approach too.

5.1 Bayesian approach to missing data

As mentioned above, a Bayesian statistician treats model parameters as unobserved random variables rather than fixed values. If one is willing to express his beliefs regarding model parameters θ as a *prior* probability distribution, $p(\theta)$, then it becomes possible to let the data update this *a priori* belief according to a result known as the Bayes' theorem[5]:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

Here, D stands for the set of observed random variables and θ for unobserved ones, $p(D) = \int p(D,\theta)d\theta$ being the marginal distribution of the random variables representing data, also known as the *evidence* for the model. Inferences are made from the conditional distribution of parameters of interest given the data. $p(D|\theta)$ is the likelihood function of the above section. The Bayes' theorem is derived from the chain rule of probability calculus that states that probability of events A and B can be expressed as the probability of A multiplied with the probability of B given A, that is, $P(A,B) = P(B|A)P(A)$. Exercise 2 asks the reader to derive the Bayes' theorem for simple random events.

The reader may have noticed that a Bayesian statistician treats both model parameters and missing/unobserved data the same, as random variables, and indeed there is no difference in this approach. Hence, Bayesian methods are easily extended for missing-data modeling. Often one is only interested in the parameters and the missing values represent so called ”nuisance” parameters. Then the Bayesian uses his rule to find the joint distribution of parameters and missing values given observed values, $p(\theta, D_{mis}|D_{obs})$, and computes the

[5] Formula was derived by Thomas Bayes (1701-1761) and the approach is, paradoxically, much older than the ”classical” frequentist statistics that was introduced by Ronald Fisher (1890-1962), Jerzy Neyman (1894-1981), and Egon Pearson (1895-1980), among others. Computational methods for actually using complex and fully Bayesian models have been developed only relatively recently, however, which explains why the Bayes' approach has occurred for most applied researchers only after the computable ”classic” one.

marginal distribution[6] of the parameters, $p(\theta|D_{obs}) = \int p(\theta, D_{mis}|D_{obs})dD_{mis}$. This marginal distribution is not biased by formally untenable computations. Everything progresses according to rules of probability calculus, and $p(\theta|D_{obs})$ automatically correctly reflects the degree of uncertainty (variance) in θ due to partial observations. In Multiple Imputation, we do not go all the way to a full Bayesian solution, but do make use of the Bayes' theorem so that uncertainty regarding D_{mis} is correctly reflected in any estimate of the parameters θ. In essence, we perform a marginalization over the D_{mis}.

5.2 Multiple Imputation as a half-Bayesian approach

It turns out that the marginal density function

$$p(D_{obs};\theta) = \int p(D_{com};\theta)dD_{mis}$$

is a valid likelihood function (for MAR and MCAR data), but not a valid sampling distribution for the true data (except for MCAR data). This explains the success of Maximum Likelihood methods for missing-data modeling. Neither Maximum Likelihood nor Bayesian methods allows us to use the familiar complete-data methods without modifying the methods themselves. This leads to many special cases that can be off-putting for an applied scientist. Hence, we would like to draw from the correct sampling distribution of D_{mis}, fill in the complete data, and do the statistical analyses we are used to doing. Figure 3 shows different ways to impute Y_{mis} in our simulated example case. One sees that simply imputing the mean of Y_{obs} miss-represents both bivariate relationship and any variability in Y_{mis}, whereas using a regression prediction (i.e., conditional mean $\mathrm{E}[Y|X]$ estimated from D_{obs}) better captures the relationship but not the variability. In Figure 3c, we sample from the regression model we have estimated from listwise-deleted data, adding also the residual variability: whatever bias in estimates of $\hat{\beta}$ and $\hat{\sigma}$ occurs due to the listwise deletion is transferred onto the sampling distribution. Furthermore, we are assuming that we have the true population parameters β and σ, even though we have only the estimates $\hat{\beta}$ and $\hat{\sigma}$ that are known to vary as a function of samples. Hence, the sampling distribution is incorrect, which shows here as a wrong slope among the imputed values.

In order to draw from the correct sampling distribution, we must take into account that we cannot gain perfect knowledge about the true population parameters based on the listwise-deleted data. The Bayes' theorem, however, allows one to estimate the posterior distribution of the parameters given the listwise deleted data, $p(\theta|D_{obs})$, that does take the uncertainty about 'true' θ into account[7]. Hence, we *can* sample values from the *joint distribution*

$$p(\theta, D_{mis}|D_{obs}) = p(D_{mis}|\theta, D_{obs})p(\theta|D_{obs}).$$

In practice, we draw a random vector θ^* from $p(\theta|D_{obs})$ and use it as our 'true' population parameter, and then draw another random vector from $p(D_{mis}|\theta, D_{obs})$ and append it to D_{obs} to complete the dataset. But now the completed dataset

[6] This would be a good time to remind oneself regarding what is "marginal distribution": http://en.wikipedia.org/wiki/Marginal_distribution

[7] Notice that here we made use of the MAR assumption of equation 2

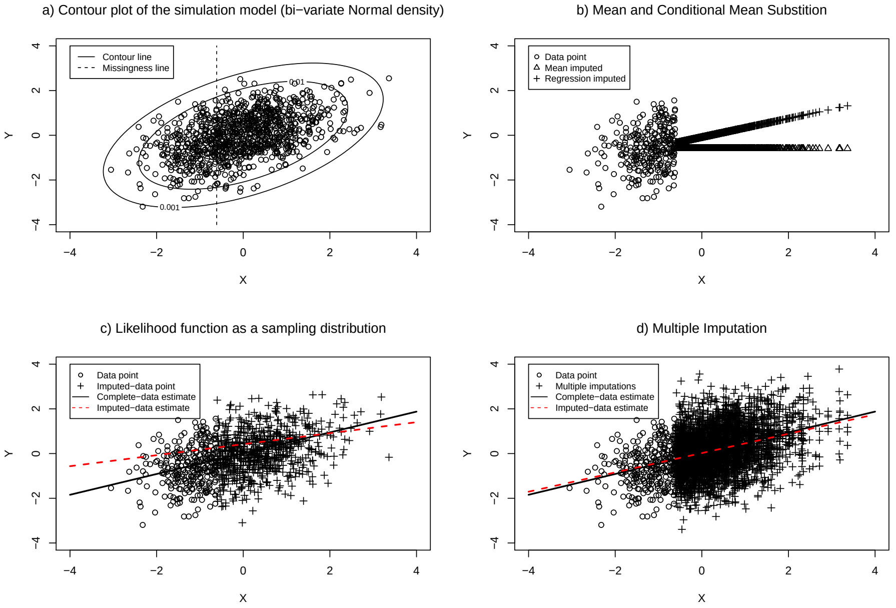


Figure 3: Explicit imputations. *a*) The data was simulated from a bivariate Normal distribution with mean (0, 0) and variances of 1, and covariance of 0.5. Contour lines of the data distribution are shown by the solid lines, whereas the dashed line shows the quantile 0.25 of X. Values exceeding that threshold in X were set as missing in Y, creating a data set with missing values missing at random (MAR). *b*) Values imputed as the mean and conditional mean (regression estimate). *c*) Values imputed from a full regression model, including residual variance. Both regression estimate based on the 'true' complete data (solid black line) and on the imputed values (dashed red line) are shown. *d)* Multiple proper imputations ($m = 5$) including both residual variance and parameter uncertainty. Both regression estimate based on the 'true' complete data (solid black line) and on an average of estimates from the imputed values (dashed red line) are shown.

we have achieved is explicitly associated with a random value θ^*, and so are the familiar estimates we compute from it, such as the regression coefficient. Let us denote the standard estimate of interest associated with $\theta^{(t)}$ as $\hat{\beta}^{(t)}$. Then, the Multiple Imputation method solves the issue by computing a Monte Carlo estimate for the expected value

$$\mathrm{E}_{(\theta,D_{mis}|D_{obs})}[\hat{\beta}] \approx \frac{1}{m}\sum_{t=1}^{m}\hat{\beta}^{(t)}$$

That is, the Multiple Imputation estimate of a model parameter, such as the regression coefficient, is an average of the estimates in m complete datasets that properly reflect the variability/uncertainty regarding the parameters needed for sampling the missing values. Hence, a Bayesian engine is in the heart of the Multiple Imputation, but the classic estimates are the end result. Monte Carlo integration refers to the fact that if we can sample random values $z^{(t)}$ from a distribution P of a random variable Z, then $\frac{1}{m}\sum_{t=1}^{m} g(z^{(t)})$ is an increasingly accurate estimate of the integral/expectation $\int g(z)p(z)dz$ as m grows. In Multiple Imputation, the recommended value of m grows with the fraction of missing data, although a surprisingly low value of m often suffices (e.g., $m = 5$). General rule of thumb is that m should be at least equal to the percentage of incomplete cases (White, Royston, & Wood, 2011). Figure 3d shows the improvement in the estimate based on imputed values from multiple proper imputations.

In summary, Multiple Imputation (*i*) constructs m completed datasets that all have slightly different values for D_{mis} but the same ones for D_{obs}, (*ii*) estimates the classic statistical model in *all* of the datasets, and then (*iii*) combines the estimates using Rubin's Monte-Carlo rules to a single unbiased estimate that was originally wished for but not obtained due to nonresponses in the data. The mentioned Rubin's rules are the following.

5.3 Rubin's rules

Multiple Imputation method creates multiple completed datasets that can be analyzed using the same methods as one would use for a dataset with no missing values. The problem is then that we need to combine the multiple results in multiple datasets to a single result that can be interpreted. Consider that we are interested in a quantity θ. As dicussed above, the way to combine the multiple estimates $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ to the single $\hat{\theta}$ is the simple averaging

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j.$$

Typically, we are not just interested in the value of the estimate $\hat{\theta}$, but also about its variance. For the jth imputed data set, we get an estimated variance of $\hat{\theta}_j$, denoted here by W_j, using standard methods, but this does not reflect the uncertainty about the true values of imputed observations that can only be seen in the differences among the imputed datasets. Therefore, the total variance of $\hat{\theta}$ is computed from the within-imputation variance $W = (1/m)\sum_{j=1}^{m} W_j$ and the between-imputation variance $B = (1/(m-1))\sum_{j=1}^{m}(\hat{\theta}_j - \hat{\theta})^2$ as the weighted sum:

$$\mathrm{Var}(\hat{\theta}) = W + \left(1 + \frac{1}{m}\right)B.$$

Standard Wald-type confidence intervals and statistical tests for $\hat{\theta}$ can be computed based on a Student's t approximation

$$\frac{(\hat{\theta} - \theta)}{\sqrt{\mathrm{Var}(\hat{\theta})}} \sim t_{\nu},$$

where the degrees of freedom are given by

$$\nu = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2.$$

These are the classic rules for combining multiple imputation estimates that were given by Rubin (1987), but better performing approximations for large-sample statistical tests have been presented since (Li, Raghunathan, & Rubin, 1991). Exercise 3 is about using Rubin's rule to compute pooled regression coefficients from those estimated in multiple imputations of Figure 3d.

6 Practical issues

6.1 MNAR data

Using the Bayesian Multiple Imputation methods or Maximum Likelihood methods, one may achieve a good performance in parameter estimation without knowing the specific mechanism for missingness, provided that it is MAR. In other words, it suffices that we include to the imputation analysis a variable(s) that predicts the nonresponse in the variable(s) with missing values. In our simulation example, the variable Y was missing for the participants with $-0.613 < X$, but all values of the X were observed. Good performance was achieved by the EM algorithm and Multiple Imputation even though we did not explicitly provide the methods with the information that the missing-value mechanism was based on the rule: "$-0.613 < X \rightarrow$ nonresponse in Y". MAR is also called *ignorable* missing data, for this reason: we can ignore the specific mechanism that gives rise to the missingness. Had the rule been "$-0.613 < Y \rightarrow$ nonresponse in Y", the missingness had not been at random (i.e., MNAR). In this case, we obviously cannot include another variable predicting the missingness, because it does not depend on anything but the variable Y itself. Tailoring methods for this kinds of special cases is possible, but difficult and the methods remain non-general. MNAR is hence also known as *nonignorable* missing data. If one stubbornly applies the methods of Figure 2 in this MNAR example (ignores the nonignorable), all estimates are biased (see Figure 4).

In the messy real-world examples, the data may be a mixture of MAR and MNAR cases, and is unlikely to have strict inequalities as missingness mechanisms. Because one rarely has a good intuition about the true missingness mechanism, it is a good idea to do at least one sensitivity analysis using an alternative missing-data approach for the important results. Notice also that our example for the missingness mechanism is rather extreme, and many real-world cases are likely to be much closer to MCAR situation. On the other hand, we have assumed linear/normal models all the time, and real data may not fit to ones' statistical model in the first place, causing bias to all estimates in general,

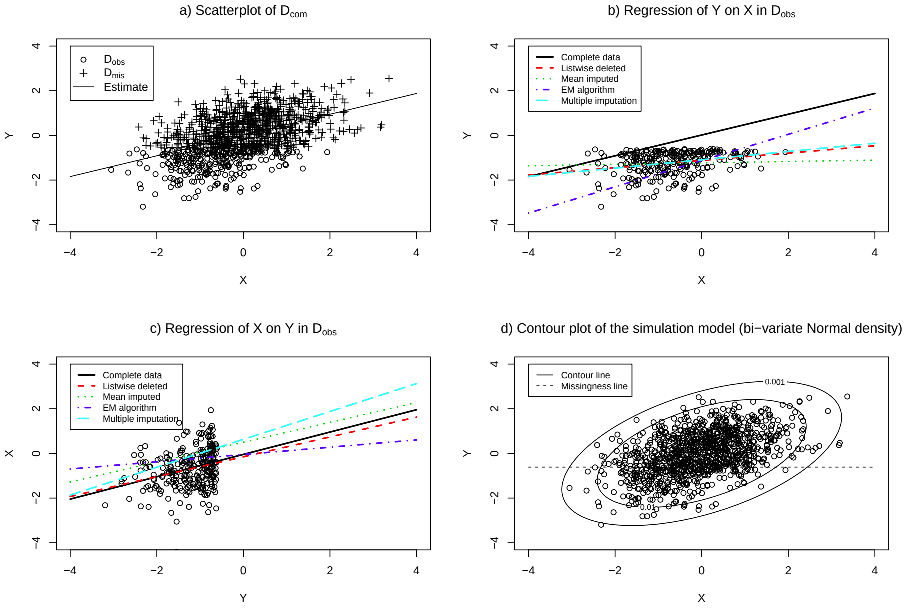


Figure 4: Results for methods of Figure 2 applied to data missing not at random (MNAR data). In Figure 2 the missingness mechanism was "$-0.613 < X \rightarrow$ nonresponse in Y", resulting in data missing at random (MAR). In this figure, the simulation was otherwise exactly the same, but the missingness mechanism was "$-0.613 < Y \rightarrow$ nonresponse in Y", resulting in MNAR data.

and specifically to imputations from that model. Hence, it is recommended to do diagnostic checks for both model fit and missingness mechanism.

For a hypothetical example, consider doing a study regarding the association between depression score, Y, for hospital nurses and their recent report regarding degree of harsh job demands in their work sector, X. Assume that those higher in depression are more likely to not respond to the depression questionnairy, as often happens, leaving us researchers with a nonresponse in Y. If the nonresponse is actually caused by the current job situation/demands that we have a report on, the data is MAR and we can achieve good imputation estimates. If the nonresponse is caused by intrinsic depression that perhaps also happen to predict reports of job demands, then the data is MNAR and we are likely to get biased imputation estimates with all the methods discussed herein. The Exercise 4 asks the reader to ponder which one, MAR or MNAR, would hold were we to perform this study.

6.2 Multivariate missingness patterns and chained equations

When several variables have missing values, we need to cycle through the variables imputing one in turn. Several cycles are needed to stabilize the results, and one needs to carefully monitor the convergence. Tools for this are included in many statistical packages for Multiple Imputation using Chained Equations (MICE) (e.g. van Buuren & Groothuis-Oudshoorn, 2011; Su, Gelman, Hill, & Yajima, 2011). With MICE one can model different data distribution, making it a very flexible method. Perhaps we shall continue from this later on, but for now, we just mention that the core ideas of MICE are similar to what has been discussed herein, with the exception that convergence issues of cycles can introduce errors and require caution. When MICE is used, the Rubin's rules can be directly applied to statistics like mean, proportion, regression coefficient, linear predictor, C-index, area under the ROC curve. Statistics that may require further transformations include odds ratio, hazard ratio, baseline hazard, survival probability, standard deviation, correlation, proportion of variance explained, skewness, and kurtosis. Rubin's rules cannot be used to combine P-value, likelihood ratio test statistic, model chi-squared statistic, or goodness-of-fit test statistic (White, Royston, & Wood, 2011).

References

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1-38.

Enders, C.K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6, 352-370.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large-sample significance levels from Multiply Imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 416, 1065-1073.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models.* John Wiley and Sons, Inc., New York, NY-US.

Muthen, B., Kaplan, D., and Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431-462.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, Inc., New York, NY-US.

Schafer, J. R. and Graham, J. W. (2002). Missing data: our view of the state of art. *Psychological Methods*, 7, 147-177.

Su, Y., Gelman, A., Hill, J., and Yajima, M. (2011). Multiple Imputation with diagnostics (mi) in R: opening windows into the black box. *Journal of Statistical Software*, 45(3), 1-67.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(2), 1-31.

White, I.R., Royston, P. and Wood, A M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377-399.

Exercises

1. Explain why listwise deletion works better in the situation of Figure 2b than in the 2c. *Hint:* Recall that estimate of a simple regression coefficient β in a model with X predicting Y is of the form

$$\beta = \mathrm{Cov}(X, Y)/\mathrm{Var}(X) = \mathrm{Cor}(X, Y)\sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}},$$

where $\mathrm{Cov}(X, Y)$ stands for the covariance of X and Y, Cor for correlation, and Var for variance.

2. Derive the Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

for two random events, A and B, from the chain rule $P(A, B) = P(B|A)P(A)$ using basic algebra.

3. When I imputed five data sets for the Figure 4c, I got the following regression-slope estimates in the individual imputed sets

i	$\hat{\beta}_i$	$\mathrm{Var}(\hat{\beta}_i)$
1	0.458	0.045
2	0.228	0.045
3	0.466	0.049
4	0.537	0.047
5	0.472	0.047

Compute the pooled estimate and its variance using Rubin's rules. If time permits, test the null hypothesis: $\theta = 0$.

4. Consider studying the association between hospital nurses' depression and job demands, and having complete observations about job demands but some degree of nonresponse in the depression scores. Is the data MAR or MNAR? Can we say anything about this question? *Hint:* Search for published literature regarding causality between nurses' depression and job demands.